# A Model For Phylogenetic Inference Based On DNA Sequences

by

Hernández, C.M.

and

Crossa, J.

BU-1212-M

June 1993

# A MODEL FOR PHYLOGENETIC INFERENCE BASED ON DNA SEQUENCES

C.M. Hernández[1] and J. Crossa[2]

Keywords: Phylogeny, DNA, convergent evolution, evolutionary tree.
Category: Statistical genetics, Modeling.

[1] Universidad de Colima, Apdo. Postal 36, Tecomán, Colima, México.
Present Address: Biometrics Unit, 336 Warren Hall, Cornell University, Ithaca, New York, 14853, U.S.A.
[2] Biometrics and Statistic Unit, International Center for Maize and Wheat Improvement (CIMMYT), Apdo. Postal 6-641, 06600 México, D.F., México.

## SUMMARY

A methodology to construct a phylogenetic tree with an associated level of confidence is proposed. It is based in modeling the number of segregating sites between two sequences after a time t from divergence, allowing for the possibility of convergent evolution. The methodology is illustrated with an example using data from Brown et al (1982).

Evolution of a particular species consists of changes in the constitution of its DNA. The genetic basis of evolution are mutations in the basis of the DNA, which can be insertions, deletions or substitutions of one of the four basis Adenine (A), Tymine (T), Cytosine (C) and Guanine (G). When comparing DNA sequences among species it is possible to infer evolutionary relationships from them. However, this is a complicate process due to evidence of "substitution preference" and different rate of mutation of some sites (Weir, 1990).

Weir (1990) presented a review and description of the three principal methods for constructing phylogenetic trees: cluster analysis, parsimony, and maximum likelihood.

The cluster method uses a distance matrix where every cell of the matrix is a measure of dissimilarity between each pair of these sequences. The problem with this method is that different distance measurements can be used and different cluster strategies can be applied so that different results can be obtained. Weir (1990) pointed out that the cluster methodology is appropriate when the mutation rates were the same on the given branches of the tree.

For a given phylogeny, the parsimony method determines the smallest number of nucleotide substitutions that will explain the observed phylogeny. The most parsimonious phylogeny is the one with the fewest number of mutations. Felsenstein (1983) pointed out that although the parsimony method generates advanced combinatorial optimization problems it is not based, like maximum likelihood, on probabilistics models. The maximum likelihood method, on the other hand, finds the segment lengths of a given tree that maximize the likelihood function. The likelihood function is the product of probabilities of independent mutations occurring in different sites of the DNA sequence.

Most of the usual procedures to construct phylogenies are time consuming and do not provide a level of certainty for the final tree(s), i.e., although it is possible to know which is the most likely tree, it is impossible to know how likely this tree is.

In this study we propose a method for phylogeny

reconstruction that is based on modelling the number of segregating sites between two sequences after a time t of divergence. The proposed model considers the possibility of convergent evolution. We use this model to construct a phylogenetic tree for three species, and then generalize the procedure to any number of species. A hypothesis test for approximating the probability of a tree to be true is also proposed.

## METHODOLOGY

The primary objective is to model the number of segregating sites between two sequences after a time t. The model's assumptions are:

1. There are K matched sequences of a Jukes-Cantor (1969) process, in which all nucleotides have the same probability to mutate to any of the remaining three bases.
2. The time between two mutations in any sequence of size N follows the exponential distribution with parameter $N\mu_o$. (No estimators of $\mu_o$ are required).
3. Nucleotides at different sites of DNA have evolved independently, that is, they have the same probability of mutation 1/N.

Since the time between two successive mutations in any sequence follows the exponential distribution with parameter $N\mu_o$, the number of mutations in a sequence A over a fixed period of time $\underline{t}$ namely $X_A(t)$, is Poisson distributed with parameter $N\mu_o t$. When considering mutations in two DNA sequences A and B, the time between two successive mutations follows the exponential distribution with parameter $2N\mu_o$. Therefore, the probability mass function of the number of substitutions in both sequences (A and B) of size N over a period of time t, namely $X_{AB}(t)$, is Poisson distributed with parameter $2N\mu_o t$

$$P(X_{AB}=x)=\frac{e^{-2N\mu_0 t}(2N\mu_0 t)^x}{x!}$$

At this point, we can consider that $N\mu_o$ is the "overall" mutation

rate of the sequence, we can make $\mu=N\mu_o$. Now, $\mu$ is measured in number of mutations per unit of time. As our objective is to construct a tree without time scale, we can change the time scale so that the number of mutations is one per unit of time. Hereafter t is expressed in this new scale. Then, the above density function

can be rewritten as:

$$P(X_{AB}=x)=\frac{e^{-2t}(2t)^x}{x!}$$

## Probability density function of the number of different nucleotides in two DNA sequences

Consider the site $i^{th}$ in a pair of DNA sequences (A and B) and call it "even site" if both nucleotides are equal and "odd site" if not. We will assume that at time of divergence t=0, every site is an even site. After a period of time t, the number of mutations that occured in a given site of two sequences may well have changed this pair so that it is no more an even site, or it could happen that this site is an even site due to convergent evolution.

Let g(t) be the probability that site $i^{th}$ is odd at time t. Then, P(site $i^{th}$ is odd) = g(t) and P(site $i^{th}$ is even) = 1-g(t). Upon defining a success when we get an odd site after a time t, we are interested in the number of successes in N independent Bernoulli trials, i.e. $J_{AB}$. Then $J_{AB}$ follows the binomial distribution with parameters N and g(t)

$$P(J_{AB}=j)\binom{N}{j}[g(t)]^j[1-g(t)]^{N-j} \qquad (1)$$

## Calculation of g(t)

Given that the number of polymorphic sites for two sequences is $J_{AB}$ ($J_{AB}$=0,1,2,3..,N), the next mutation will make this number to be $J_{AB}$-1, $J_{AB}$ or $J_{AB}$+1, with the following probabilities:

$P(J_{AB}=J_{AB}-1)$=P(segregating sites decreases in 1) = J/3N,

$P(J_{AB}=J_{AB})$=P(segregating sites does not change) = 2J/3N,

$P(J_{AB}=J_{AB}+1)$=P(segregating sites increases in 1) = (N-J)/N.

In the above set of transition probabilities, $P(J_{AB}=J_{AB}-1)$ and $P(J_{AB}=J_{AB})$ necessarily imply that an event exists in which mutations occur more than once in a given site. This is a Markov Chain process with the following transition probability matrix:

$$P=\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1/3N & 2/3N & (N-1)/N & 0 & 0 & \cdots & 0 & 0 \\ 0 & 2/3N & 4/3N & (N-2)/N & 0 & \cdots & 0 & 0 \\ 0 & 0 & 3/3N & 6/3N & (N-3)/N & \cdots & 0 & 0 \\ 0 & 0 & 0 & 4/3N & 8/3N & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & & 2(N-1)/N & 1/N \\ 0 & 0 & 0 & 0 & \cdots & & 1/3 & 2/3 \end{pmatrix}$$

Thus, the transition probabilities are:

$$\beta_t=\frac{N-i}{j} \qquad (j=i+1),$$

$$\delta_i=\frac{i}{3N} \qquad (j=i-1),$$

$$\alpha_i=\frac{2i}{3N} \qquad (j=i), \text{ and}$$

0 otherwise

Now we find the expected value and the variance of $J_{AB}$ after time t from divergence and after n mutations.

## 1. Expectation and variance of $J_{AB}$ after n mutations.

Let $e_n = E\{J_n\}$ be the expected value of $J$ after a total of n mutations in both sequences A and B. We have:

$$\begin{aligned} e_n = E(J_n) &= E(J_{n-1} + J_n - J_{n-1}) \\ &= E(J_{n-1}) + E(J_n - J_{n-1}) \\ &= e_{n-1} + E(J_n - J_{n-1}) \end{aligned}$$

To evaluate $E(J_n - J_{n-1})$ first we calculate the conditional expected value of $J_n - J_{n-1}$ given $J_{n-1}$ and then take the expected value of this conditional mean. First we have that given $J_{n-1}$, the difference $J_n - J_{n-1}$ has the following mass function:

$$f(J_n - J_{n-1}) = \begin{cases} J_{n-1}/3N, & \text{if } J_n = J_{n-1}-1 \\ 2J_{n-1}/3N, & \text{if } J_n = J_{n-1} \\ (N - J_{n-1})/N, & \text{if } J_n = J_{n-1}+1 \end{cases}$$

The expectation of $J_n - J_{n-1}$ for a fixed $J_{n-1}$ is:

$$1-(4J_{n-1})/3N$$

and the expectation of this conditional mean is

$$1-(4e_{n-1})/3N$$

since

$$e_n = e_{n-1} + E_i(J_n - J_{n-1})$$

we have

$$e_n = e_{n-1} + 1-(4e_{n-1})/3N$$

$$= 1 + e_{n-1}(1-4/3N)$$

Using this relationship successively n-1 times, it can be shown that the expected value for the number of segregating sites for sequences A and B after n mutations is

$$E(J_n) = e_n = \frac{1-[1-(\frac{4}{3})]^n}{4/3N} \qquad \text{(Appendix A)} \qquad (2)$$

Using a similar argument it can be shown that

$$e^2{}_n = (\frac{1-K^n}{1-K})(1+\frac{3NQ}{2}) - (\frac{3NQW^{n-1}}{2})(\frac{1-(K/W)^n}{1-K/W}) \qquad (3)$$

(Appendix B) where

W = 1-4/3N, Q = 1-1/3N, and K = 1-8/3N.

The variance of $J_n$ given n mutations is

$$V(J_n) = e^2{}_n - (e_n)^2 \qquad (4)$$

It can be easily verified that when $n^2 \to \infty$

$$E\{J\} = 3N/4 \qquad \text{and}$$
$$V\{J\} = 3N/16$$

Using the results above, we can find the expectation and the variance of J after a time t from divergence.

## 2. The expected value and variance of $J_{AB}$ after a time t from divergence

We use the p.m.f. of the number of mutation substitutions (n) in both sequences (A and B) which is, as shown before, Poisson with parameter 2t. After a period of time t, n can take values n=0,1,2,... If we denote E{J/n=i} as the expected value of J given that the number of mutations in both sequences have been i, the expected value of J after a time t is, by the law of total probability:

$E\{J|\ T=t) = E\{J/n=0\}P(n=0)+E\{J/n=1\}P(n=1)+E\{J/n=2)P(n=2)+...$

$$E(J/T=t) = \sum_{n=0}^{\infty} ((\frac{1-(1-4/3N)^n}{4/3N})(e^{-2t}\frac{(2t)^n}{n!}))$$

which after simplification gives

$$E\{J|T=t\}=\frac{3N}{4}(1-e^{-8t/3N}) \quad \text{(Appendix C)} \qquad (5)$$

Also the variance of J is

$$V\{J|T=t\}=(\frac{3N}{16})(2e^{-8t/3N}-3e^{-16t/3N}+1) \qquad (6)$$

as we would expect for a binomial distribution.

Then, the value of g(t) in expression (1) is:

$$g(t)=\frac{3}{4}(1-e^{-8t/3N})$$

Thus, the density function of $J_{AB}$ is:

$$P(J_{AB}=j)=\binom{N}{j}(\frac{3}{4}(1-e^{-8t/3N}))^{j}(1-\frac{3}{4}(1-e^{-8t/3N}))^{N-j}$$

The steady - state distribution of $J_{AB}$, namely $\pi_j$, when $t \to \infty$ is

$$\pi_j=\binom{N}{j}(3/4)^{j}(1/4)^{N-j} \quad 0\le j\le N \qquad (7)$$

## Hypothesis Testing

### Three species

The three random variables that are obtained when we have three DNA sequences, namely A, B and C, are: 1) $J_{AB}$ the number of segregating sites between sequences A and B, 2) $J_{AC}$ the number of segregating sites between sequences A and C, and 3) $J_{BC}$ the number of segregating sites between sequences B and C. It is important to make some remarks on this data: i) it is not necessary that the data came from contiguos sites, indeed, it could be a sample of size N, ii) they must come from aligned sequences, that is, site i is the same for all sequences. This implies that the information

provided by sites were there has been insertions and/or deletions is not considered.

If for a given test, it is found that the probability that the three species have the same time to divergence is very small, then the only alternative is to consider a tree like the depicted in Fig. 1.1. However, if $J_{AC}$, $J_{AB}$ and $J_{BC}$ are very alike, we have to conclude that there is not enough information where to place the 3 species in the branches of the tree and Fig. 1.2 is the only conclusion. This does not imply that the three species have the same time to divergence, but that the tree can not be solved, we adopt the term "null tree" for Fig. 1.2

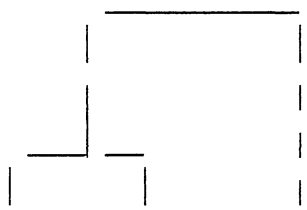


Fig. 1.1

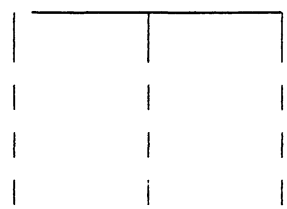


Fig. 1.2

It is reasonable to assume that those species placed in the shorter branches will have the smallest value for the number of segregating sites (J), since they have shorter time to a common ancestor. Let $t_o$ be the time to divergence of the three species and $t_1$ be the time of divergence of A and C (Fig. 3).

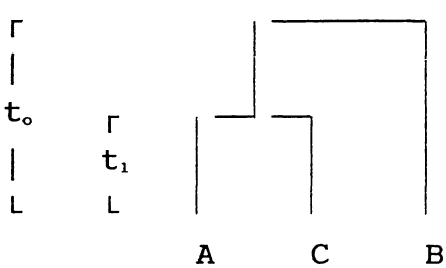



Fig. 3

Since the time to divergence of A and B is equal to that of C and B, we have that $E\{J_{AB}\} = E\{J_{CB}\}$, we also have that $E\{J_{AC}\} \leq$

[$E\{J_{CB}] = E\{J_{AB}\}$]. Then, if $t_o > t_1$ we will conclude that the appropriate three is Fig. 1.1, and if $t_1 = t_o$ we conclude the null tree. The set of hypothesis is:

$$H_o: t_o = t_1$$
$$H_1: t_o > t_1$$

## The probability density function of a function of three random variables ($J_{AB}$, $J_{AC}$, and $J_{BC}$) under the null hypothesis

First we shall find an expression for the covariance for any pair of the three r.v.'s, $J_{AB}$, $J_{AC}$, and $J_{BC}$ when the three species have the same time to divergence. We know that

$$Cov(\sum_{i=1}^{N} X_i, \sum_{j=i}^{N} Y_j) = \sum_{i=1}^{N}\sum_{j=1}^{N} (Cov(X_i, Y_j))$$

If we let $X_i$ be 0 if site i in sequences A and C is an even site and 1 if do not, we have that $J_{AC}$ is the sum of these independent random variables. Similarly $Y_j$=0 if site j in sequences A and B is an even site and 1 if do not.

By independence $Cov(X_i,Y_j)=0$ $(i \neq j)$ therefore

$$\sum_{i=1}^{N}\sum_{j=1}^{N} (Cov(X_i, Y_j)) = \sum_{i=1}^{N} Cov(X_i, Y_i) = N(Cov(X, Y))$$

where X and Y are the values of the random variables for pairs AC and AB in the same site. Now we consider the product Z=XY. Z can take only two values: 0 or 1, therefore the expected value of Z after a time t of divergence of the three species is the probability that Z takes the value 1, that is, the probability that X=1 an Y=1 after a time t.

It is possible to show that given "a" substitutions in a given site in sequence A and "c" substitutions in the same site in sequence C, the probability that X takes the value 1 is given by $\frac{3}{4}[1-(-\frac{1}{3})^{a+c}]$ **(Appendix D)**. When calculating the expectation of $J_{AC}$, the probability that X and Y take the value 1 after a time t can be

evaluated by considering that given that X=1, the probability that Y=1 depends only on the number of mutations occurred in sequences A and B. Thus we have

$$\mathrm{P}(X=1, Y=1) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\sum_{c=0}^{\infty} \mathrm{P}(a)\,\mathrm{P}(b)\,\mathrm{P}(c)\,\mathrm{P}(X=1 \mid a+c)\,\mathrm{P}(Y=1 \mid a+b)$$

where P(a), P(b) and P(c) are the probabilities of exactly a, b, and c mutations in sequences A, B, and C, respectively. The latter expression can be rewritten as:

$$P(z=1) = P(X=1, Y=1) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\sum_{c=0}^{\infty} P(a)\,P(b)\,P(c)\,\left(\frac{3}{4}\right)^2\left[1-\left(-\frac{1}{3}\right)^{a+c}\right]\left[1-\left(-\frac{1}{3}\right)^{a+b}\right]$$

As previously shown, the density function of the number of substitutions in only one sequence of size N is Poisson with parameter t. Thus we have that the parameter for one single nucleotide is (t/N) and therefore

$$\mathrm{P}(a=r) = \frac{e^{-t/N}(t/N)^r}{r!}$$

Using this expression, we can show that:

$$E(Z) = E(XY) = P(X=1, Y=1) = \frac{9}{16}\left[1-2e^{-8t/3N}+e^{-32t/9N}\right]$$

**(Appendix E)** therefore

$$Cov(J_{AC}, J_{AB}) = \frac{9N}{16}\left[e^{-32t/9N}-e^{-16t/3N}\right] \qquad (8)$$

If we are willing to assume that the pair of species with shorter time from divergence (if any) will have smaller J-value, then we are interested in testing if this value is statistically smaller than the other two, which then come from the same population. A test based in this assumption is reasonable and tends to reduce the probability of concluding a wrong tree, as it will be shown later. Without loss of generality we let A and C be the species with shorter time from divergence $t_1$, whereas the pairs AB and BC have larger time from divergence $t_0$. Thus, $J_{AC}$ is a sample

from a binomial population $(N,g(t_1))$ whereas $J_{BC}$ and $J_{AB}$ are two samples from $(N,g(t_0))$, $t_1<t_0$.

Instead of looking at Types I and II error probabilities, we look at the more general probability of concluding a wrong tree. The importance of evaluating this probability is that we still can make an error if $H_0$ is false and it is rejected, if it happens that we do not place correctly the species on the branches. If $H_0$ is true, then P(wrong tree) reduces to $\alpha$ for a level $\alpha$ test. On the other hand, if $H_1$ is true, then P(wrong tree) is the probability that $H_0$ is rejected and $J_{AC}$ is not the minimum value of the J's, in which case we do not place the species A and C in the shortest branches of the tree. Note that we are not concerned with the probability of a Type II error since if $H_0$ is not rejected the conclusion is that the tree can not be solved with the information given.

A level $\alpha$ test can be implemented as follows: let Cov be the value of the covariance and $\sigma^2$ the variance of any of the J's. Under the null hypothesis, the distribution of the random variable defined as:

$$J_{\bar{x}}=\frac{J_{AB}+J_{BC}}{2}$$

is normal with mean $E\{J_{AC}\}$ and variance $(\sigma^2+Cov)/2$. Also,

$$J_x - J_{AC}$$

is normal with mean 0 and variance $3(\sigma^2-Cov)/2$.

Under $H_0$, the difference between $J_x$ and $J_{AC}$ should be small, and the statistic d defined as:

$$d=\frac{J_{\bar{x}}-J_{AC}}{\sqrt{3\,(\sigma^2-Cov)\,/2}} \qquad (9)$$

is normal standard, and it can be used to test the set of hypothesis. **(Appendix F)**

We reject $H_0$ if $d \geq Z_\alpha$. It can be shown that for this test, the following inequality is true:

$$P(\text{Wrong tree} \mid H_1 \text{ true}) < P(\text{Wrong tree} \mid H_0 \text{ true}) < \alpha$$

**(Appendix G)**

The estimator of $t_0$ can be constructed by considering that under $H_0$ the three J's have the same expected value, therefore:

$$E\left(\frac{J_{AC}+J_{AB}+J_{BC}}{3}\right) \mid T=t_0) = Ng(t_0)$$

$$Ng(t_0) = \frac{3N}{4}(1-e^{-8t_0/3N}) \quad \text{and}$$

$$t_0 = \left(\frac{-3N}{8}\right)\left(\ln\left(1-\frac{4\bar{x}}{3N}\right)\right) \tag{10}$$

where X is the mean of the three r.v.'s $J_{AC}$, $J_{AB}$, and $J_{CB}$.

**EXAMPLE**

We illustrate the methodology using data from Brown et al (1982) which consists of 896 bases of mitochondrial DNA sequences of five primates: Human (H), Chimpanzee (Ch), Gorilla (Go), Orangutan (Or) and Gibbon (Gi). The J values for every pair of species is shown in Table 1.

Table 1. J values for the five primates

| | H | Ch | Go | Or | Gi |
|---|---|---|---|---|---|
| H | | 79 | 92 | 144 | 169 |
| Ch | | | 95 | 154 | 169 |
| Go | | | | 150 | 169 |
| Or | | | | | 169 |

First we take H, Ch and Go to show the method for only 3 species. We have that $J_{HCh}=79$, $J_{HGo}=92$, and $J_{ChGo}=95$. With this data we

calculate $t_0$ = 47.543, $\sigma^2$ = 79.892, Cov = 37.569, and d = 1.819. We reject $H_0$ with an $\alpha$ level of 0.043. The smallest value of J is the corresponding to $J_{HCh}$, therefore, we conclude that Human and Chimpanzee are the species with shorter time to divergence. We introduce the terminology "H and Ch have a unitary relationship with Go" which will be used later. The resulting tree is depicted in Fig. 4 and has $1-\alpha$= 1-0.0351= 0.9649 probability of express the true relationship between these species:

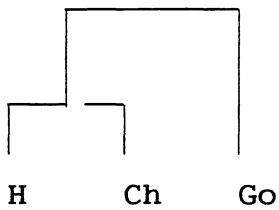
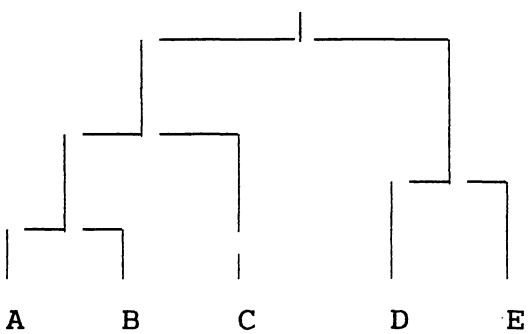



Fig. 4

**More than three species**

Consider the following arbitrary tree:




Note that the time from divergence (and hence, the expected number of segregating sites) is the same for AC and BC, also the time from divergence for pair AD is the same as for pairs AE, BD, BE, CD, and CE.

The strategy for more than three species starts by testing the pair of species with smallest value of J. Every time it is found that a given pair of species has unitary relationship, we will not longer consider them as two separate species, instead, we consider

this set as a fictitious new specie and the J-value for this specie with another one will be the mean of the Js. The first test always involves only three species, and the result is the fusion of the pair with smallest J-value or the fusion of all three, depending upon $H_o$ being rejected or not. Nevertheless, as the number of species added increases, it well could happen that the test involves three groups of species, depending on the topology of the tree.

To simplify our procedure, we will always consider finding the phylogeny for three species $S_A$, $S_B$ and $S_C$, where the pair of species $S_A$ and $S_C$ have smaller J-value. It is very useful to record the result of every successive test in a matrix whose reading will provide us with the final tree. We illustrate this process using data from Brown et al. (1982)

**STEP 1.**

1) The initial J-values are those from Table 1. The smallest of the J's is for H-Ch. In the case that the minimum is not unique, we can take any of them.

2) To select the additional specie, we choose the specie which is "closest" to the pair H-Ch. This specie is Go, since $(J_{HGo}+J_{ChGo})/2 = 93$ is minimum over the species Go, Or and Gi.

3) From 1 and 2 we have that the corresponding species $S_A$, $S_B$ and $S_C$ are : $S_A$ = {H}; $S_C$= {Ch}; $S_B$ = {Go}.

4) We use data from 3 to test if $S_A$ and $S_C$ have unitary relationship with respect to $S_B$. From the previous example we have already found that $H_o$ is rejected with $\alpha = 0.0351$.

5) Those species that had unitary relationship are assigned 0 and 1. If $H_o$ is not rejected then we assign 0,1 and 2. The assignation is indistinct to the species. In this case $S_A$= 0 and $S_C$= 1, or H=0 and Ch=1.

6) We fusion species $S_A$ and $S_C$, and call this new specie T1. Now we construct table 2 as follows:

Table 2. Results of step 1.

| | T1 | Go | Or | Gi |
|---|---|---|---|---|
| T1 | | 93 | 149 | 169 |
| Go | | | 150 | 169 |
| Or | | | | 169 |

Note

$$J_{T1Go} = (J_{SAGo} + J_{SCGo})/2 = (J_{HGo} + J_{ChGo})/2 = 93$$

$$J_{T1Or} = (J_{SAOr} + J_{SCOr})/2 = (J_{HOr} + J_{ChOr})/2 = 149$$

$$J_{T1Gi} = (J_{SAGi} + J_{SCGi})/2 = (J_{HGi} + J_{ChGi})/2 = 169$$

**STEP 2.**

1) The initial J-values are those of Table 2. We take the smallest of the J's that is (T1-Go).

2) The closest specie to the pair T1-Go is Or, since $(J_{T1Or} + J_{GoOr})/2 = 149.5 \approx 149$. At this point, we shall mention that we have adopted the convention of using the integer part.

3) The corresponding species $S_A$, $S_B$ and $S_C$ are : $S_A$= {T1}; $S_C$ = {Go}; $S_B$ = {Or}.

4) With data from 3 we test if $S_A$ and $S_C$ have unitary relationship with respect to $S_B$. We use expressions (6), (8), (9) and (10) which yield: $t_o = 72.6509$; $\sigma^2 = 111.611$; Cov = 50.7118; d = 5.9114 Thus $\alpha$ is negligible for this test.

5) Since $S_A$ and $S_C$ have unitary relationship, we assign to those species 0 and 1, this means that H and Ch are assigned a 0 and Go a 1.

6) Again we fusion species $S_A$ and $S_C$, in this case T1 and Go, and call it T2. The new table is as follows:

Table 3. Results of step 2.

| | T2 | Or | Gi |
|---|---|---|---|
| T2 | | 149 | 169 |
| Or | | | 169 |

Note

$$J_{T2Or} = (J_{SAOr} + J_{SCOr})/2 = (J_{T1Or} + J_{GoOr})/2 = 149$$
$$J_{T2Gi} = (J_{SAGi} + J_{SCGi})/2 = (J_{T1Gi} + J_{GoGi})/2 = 169$$

Values for Table 3 were calculated using the values of Table 2. Although we could have used for instance, to calculate $J_{T2Or}$, the arithmetic mean of the J-values of the species that are included in T2 (H, Ch and Go) with Or. This would give similar weight to the values $J_{HOr}$, $J_{ChOr}$ and $J_{GoOr}$, which is incorrect, since H and Ch were previously found to have unitary relationship. Thus, they do not provide of independent estimations of their time to divergence with Or. In general, we only have to use the resulting table from the previous step.

**STEP 3.**

1-3) The initial J-values are those of Table 3. The smallest value is for the pair (T2-Or). There are only three remaining species, therefore: $S_A$ = {T2}; $S_C$ = {Or} and $S_B$={Gi}.

4) With data from Table 3 we test if $S_A$ and $S_C$ have unitary relationship with respect to $S_B$. We use expressions (6), (8), (9) and (10) which yield: $t_o$ = 92.9044; $\sigma^2$= 132.922; Cov = 58.6830 d = 1.8952 Thus $\alpha$ = 0.0294

5) Since T2 and Or have unitary relationship, we assign them 0 and 1. Since $S_A$ = {H, Ch, Go} we have H=0, Ch=0, Go=0, Or=1.

To draw the tree, we use the labels assigned to every species (part 5 of every step). We have:

| | Species | | | | |
|---|---|---|---|---|---|
| Label | H | Ch | Go | Or | Gi |
| Step1 | 0 | 1 | | | |
| Step2 | 0 | 0 | 1 | | |
| Step3 | 0 | 0 | 0 | 1 | |
| Extra Row | 0 | 0 | 0 | 0 | 1 |

The extra row is just O's for all species that had a label (any) in the last step and a 1 to the remaining one(s). The column for every specie is a codeword for its localization in the tree, reading from bottom to top, thus:

| Specie | Codeword |
|---|---|
| Human | 0000 |
| Chimpanzee | 0001 |
| Gorilla | 001 |
| Orangutan | 01 |
| Gibbon | 1 |

and the tree

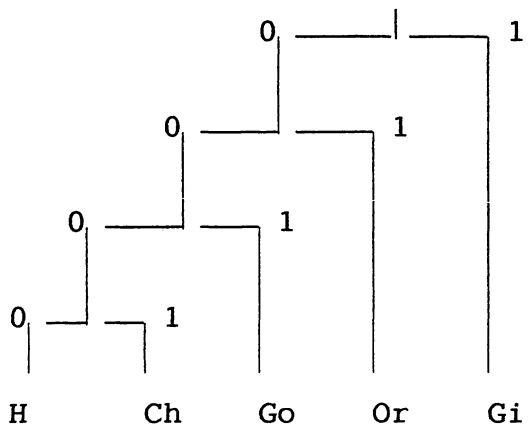
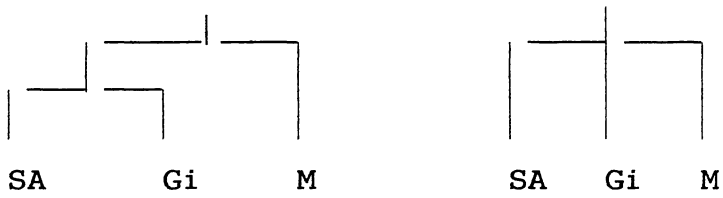



Fig. (A) Final tree for the five primates of Brown et al. (1982)

**Total confidence of the tree**

To illustrate the independence of the tests developed, we present the following argument: let P be the probability that the true philogeny for the 5 species is as depicted in fig. (A). Brown et al also published the mitochondrial sequence of Mouse (M). Its J-values with the five primates are $J_{HM}$=298, $J_{ChM}$= 296, $J_{GoM}$= 297, $J_{OrM}$= 307, $J_{GiM}$= 301. These values are so big comparatively to the primates, that if we were to consider Mouse in our tree, this specie would be the last to be added. We have $S_A$={T3}, $S_C$={Gi} and $S_B${M}. The final tree should be one of the following:




Note that the probability of adding correctly Mouse to the tree is (1-$\alpha$), were $\alpha$ is the error level for the test that involves Mouse. Thus, the total probability of the tree would be P times the probability of correctly adding Mouse to the tree, i.e. P(1-$\alpha$). This independence only arises if the species to be added is going to be placed above or at the level of the first node of the current tree.

The previous reasoning implies that we can change the J-values of Table 1, and thus, alter the topology and the associated probability of the tree, but as long as Mouse remains as the last specie that must be added, the last test is independent from the previous.

Our procedure ensures that every aded species has to be located above or at the level of the first node, therefore, every test is independent of the others, in our case, the first test gave $\alpha$=0.0351, in the second $\alpha$ is negligible, and the third $\alpha$=0.0294, thus, the probability that the final tree is true is (0.9649) (0.9706)= 0.9365

Unlike the maximum likelihood approach the proposed method allows to compare a large number of DNA sequences. However , as the

number of species involved increases, the total number of possible trees increases and therefore the confidence of the method decreases. For more than 3 species the method requires not only a large number of calculations but a continuos iterative procedure.

The authors developed a computer program that finds the phylogenetic tree for up to 100 species. The required input is the J-values for every pair of species, the size of the sequence and the $\alpha$ value to reject the null tree.

The authors wish to thank Enrique Estrada L. for his helpful comments on the manuscript, and to George Casella and Charles McCulloch for their time.

**Appendix A.**

We have

$$e_n = 1 + e_{n-1}(1 - 4/3N)$$

thus

$$e_{n-1} = 1 + e_{n-2}(1 - 4/3N)$$

let W=(1-4/3N) then

$$e_n = 1 + W + e_{n-2}W^2$$

or

$$e_n = 1 + W + W^2 + W^3 + \ldots + e_1 W^{n-1} + e_0 W^n$$

since $e_o=0$ and $e_1=1$, then we have

$$e_n = 1 + W + W^2 + \ldots + W^{n-1} = \sum_{i=0}^{n=1} W^i = \frac{1 - W^n}{1 - W}$$

then

$$e_n = \frac{1 - [1 - (\frac{4}{3})]^n}{4/3N}$$

**Appendix B.**

Letting

$$g(n-i) = 1 + 2e_{n-i}(1 - \frac{1}{3N}) \quad \text{and}$$

using successively this relationship, we have

$$e_n^2 = g(n-1) + e_{n-1}^2(1 - 8/3N)$$

let K= 1-8/3N, thus

$$e_n^2 = g(n-1) + e_{n-1}^2(K)$$

$$e_n^2 = g(n-1) + kg(n-2) + k^2 e_{n-2}^2$$

i.e.

$$e_n^2 = g(n-1) + kg(n-2) + k^2 g(n-3) + \ldots + k^{n-2} g(1) + k^{n-1} g(0)$$

since $g(0) = 1+2K\ e_o = 1$, then

$$e_n^2 = g(n-1) + kg(n-2) + k^2 g(n-3) + \ldots + k^{n-2} g(1)$$

$$e_n^2 = \sum_{i=1}^{n} g(n-i)\, k^{i-1}$$

since

$$e_{n-i} = \frac{1-(1-4/3N)^{n-i}}{4/3N}$$

let W=1-4/3N, Q=1-1/3N, and K=1-8/3N then

$$e_n^2 = \sum_{i=1}^{n} [1 + \frac{2Q(1-W^{n-i})}{4/3N}]\, k^{i-1} = \sum_{i=0}^{n-1} k^i + \frac{3NQ}{2} \sum_{i=0}^{n-1} k^i - \sum_{i=0}^{n-1} (\frac{k}{w})^i \frac{3NQ}{2} W^{n-1} =$$

$$= (\frac{1-k^n}{1-k})\ (1 + \frac{3NQ}{2}) - (\frac{3NQW^{n-1}}{2})\ (\frac{1-(k/w)^n}{1-k/w})$$

## Appendix C.

We have $E[J_{AB}/T=t] =$

$$\sum_{n=0}^{\infty} \frac{e^{-2t} 2t^n}{n!} [\frac{1-(1-4/3N)^n}{4/3N}]$$

let W=1-4/3N, then

$$E[J_{AB}/T=t] = \frac{e^{-2t}}{1-W} [\sum_{n=0}^{\infty} \frac{(2t)^n}{n!} (1-W^n)]$$

$$= \frac{e^{-2t}}{1-W}\left[\sum_{n=0}^{\infty}\frac{(2t)^n}{n!} - \sum_{n=0}^{\infty}\frac{(2tW)^n}{n!}\right]$$

$$= \frac{e^{-2t}}{1-W}\left[e^{2t} - e^{2tW}\right] = \frac{1-e^{2t(-\frac{4}{3N})}}{4/3N}$$

$$= \frac{1-e^{-8t/3N}}{4/3N} = \frac{3N}{4}\left[1-e^{-8t/3N}\right]$$

**Appendix D.**

If P is a transition matrix of the form

$$P = \begin{bmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{bmatrix}$$

then $P^n = \frac{1}{2-\alpha-\beta}\begin{bmatrix} 1-\beta+(1-\alpha)(\alpha+\beta-1)^n & (1-\alpha)[1-(\alpha+\beta-1)^n] \\ (1-\beta)[1-(\alpha+\beta-1)^n] & 1-\alpha+(1-\beta)(\alpha+\beta-1)^n \end{bmatrix}$

Now, If the site is even, the next mutation necessarily will make it an odd site, thus $P(0 \to 0) = 0$, $P(0 \to 1) = 1$. If the site is odd, the next mutation will make it an even site wih probability ⅓, thus $P(1 \to 0) = ⅓$, $P(1 \to 1) = ⅔$ and our transition matrix is

$$P = \begin{bmatrix} 0 & 1 \\ ⅓ & ⅔ \end{bmatrix}$$

so $\alpha=0$, $\beta=⅔$ and we are interested in $P(0 \to 1)$ in n steps, which is

$$\frac{(1-\alpha)\left[1-(\alpha+\beta-1)^n\right]}{2-\alpha-\beta} = \frac{1-(-1/3)^n}{4/3} = \frac{3}{4}\left[1-(-\frac{1}{3})^n\right]$$ (Tsokos, 1972)

**Appendix E.**

We have

$$P(Z=1) = P(X=1, Y=1) = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} P(a)\, P(b)\, P(c)\, P(X=1 \mid a+c)\, P(Y=1 \mid a+b)$$

As we have

$$P(a=r) = \frac{e^{-t/N}(t/N)^r}{r!} \qquad \text{and}$$

$$P(X=1 \mid a+c) = \frac{3}{4}\left[1-\left(-\frac{1}{3}\right)^{a+c}\right]$$

$$P(Z=1) = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)\left(\frac{3}{4}\right)^2$$

$$\left[1-\left(-\frac{1}{3}\right)^{a+c}\right]\left[1-\left(-\frac{1}{3}\right)^{a+b}\right]$$

$$P(z=1) = \frac{9}{16} \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)$$

$$\left[1-\left(-\frac{1}{3}\right)^{a+b}-\left(-\frac{1}{3}\right)^{a+c}+\left(-\frac{1}{3}\right)^{2a+b+c}\right]$$

let

$$A_1 = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)$$

$$A_2 = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)\left(-\frac{1}{3}\right)^{a+b}$$

$$A_3 = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)\left(-\frac{1}{3}\right)^{a+c}$$

$$A_4 = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} \left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)\left(-\frac{1}{3}\right)^{2a+b+c}$$

then

$$P(Z=1)=P(x=1,y=1)=\frac{9}{16}(A_1-A_2-A_3+A_4)$$

$A_1$ can be simplified as

$$A_1=\sum_{a=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\sum_{b=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\sum_{c=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^c}{c!}\right)=1$$

$A_2$ can be reduced as

$$A_2=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(\frac{e^{-t/N}(t/N)^b}{b!}\right)\left(-\frac{1}{3}\right)^{a+b}\sum_{c=0}^{\infty}\frac{e^{-t/N}(t/N)^c}{c!}$$

$$=\sum_{a=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^a}{a!}\right)\left(-\frac{1}{3}\right)^a\sum_{b=0}^{\infty}\frac{e^{-t/N}(t/N)^b}{b!}\left(-\frac{1}{3}\right)^b$$

$$=\sum_{a=0}^{\infty}\left(\frac{e^{-t/N}\left(-\frac{t}{3N}\right)^a}{a!}\right)\sum_{b=0}^{\infty}\left(\frac{e^{-t/N}\left(-\frac{t}{3N}\right)^b}{b!}\right)$$

since

$$\sum_{x=0}^{\infty}\frac{k^x}{x!}=e^k,\ -\infty<k<\infty$$

then we have

$$A_2=[e^{-t/N}e^{-t/3N}]^2=[e^{-4t/3N}]^2=e^{-8t/3N}$$

similarly, $A_3 = e^{-8t/3N}$

$$A_4=\sum_{a=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^a(-1/3)^{2a}}{a!}\right)\sum_{b=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^b\left(-\frac{1}{3}\right)^b}{b!}\right)\sum_{c=0}^{\infty}\left(\frac{e^{-t/N}(t/N)^c\left(-\frac{1}{3}\right)^c}{c!}\right.$$

Using previous results

$$=\sum_{a=0}^{\infty}\left(\frac{e^{-t/N}\left(\frac{t}{9N}\right)^a}{a!}\right)\sum_{b=0}^{\infty}\left(\frac{e^{-t/N}(-t/3N)^b}{b!}\right)\sum_{c=0}^{\infty}\left(\frac{e^{-t/N}(-t/3N)^c}{c!}\right)$$

$$(e^{-t/N}\ e^{t/9N})\ (e^{-t/N}\ e^{-t/3N})\ (e^{-t/N}\ e^{-t/3N})$$

$$=e^{-8t/9N}e^{-8t/3N}=e^{\frac{-8t}{9N}}e^{\frac{-24t}{9N}}=e^{\frac{-32t}{9N}}$$

Thus,

$$P(Z=1)=P(X=1,Y=1)=\frac{9}{16}(1-e^{-8t/3N}-e^{-8t/3N}+e^{-34t/9N})$$

$$=\frac{9}{16}(1-2e^{-8t/3N}+e^{-32t/9N})$$

Then Cov(X,Y)= E(XY) - E(X)E(Y)
but E(XY)= E(Z)= (O) P(Z=0)+(1) P(Z=1)= P(Z=1)= P(X=1,Y=1)
and

$$E(X)=E(Y)=g(t)=\frac{3}{4}(1-e^{-8t/3N})$$

then

$$Cov(X,Y)=\frac{9}{16}(1-2e^{-8t/3N}+e^{-32t/9N})-[\frac{3}{4}(1-e^{-8t/3N})]^2$$

$$=\frac{9}{16}(1-2e^{-8t/3N}+e^{-32t/9N})-\frac{9}{16}(1-2e^{-8t/3N}+e^{-16t/3N})$$

$$=\frac{9}{16}(e^{-32t/9N}-e^{-16t/3N})$$

## Appendix F

Under $H_0$ the J's are distributed Binomial with parameters N and g(t), but they are not independent. Altough it is possible to use normal approximation to binomial, the result $(J_{AB}+J_{BC})/2$ distributed normal is not direct due to dependence among the J's. Since every one of the random variables involved is the sum of independent bernoulli random variables, we can use the Lindberg-Levy theorem:

Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_N=\sum_{i=1}^{N} Z_i$$

then

$$\lim_{N\to\infty} P\left(\frac{S_N-N\mu}{\sigma\sqrt{N}}\le a\right)=\Phi(a)$$

for all a, $-\infty < a < \infty$, where $\Phi$ denotes the standard normal c.d.f. That is, the sum of i.i.d. can be aproximated by the normal distribution when N is large.

Let $X_i$ if site i is even for sequences A and B and 1 if not. Similarly define $Y_i$ for sequences B and C. let $Z_i=X_i+Y_i$. With $S_N$ as defined above we have:

$E\{X_i\} = E\{Y_i\} = g(t)$

$E\{Z_i\} = E\{X_i+Y_i\} = 2g(t)$

$Var\{Z_i\} = 2\ (Var\{X_i\} + Cov(X_i,Y_i))$

Since $Cov(X_i,Y_i) = Cov(J_{AB},J_{BC})/N = Cov/N$, we have:

$Var\{Z_i\} = 2[g(t)(1-g(t)) + Cov/N]$, from here:

$Z/2 =(J_{AB}+J_{BC})/2$ is normal with mean $Ng(t)$ and variance $[Ng(t)(1-g(t))+Cov]/2$. Since under $H_0$ $Ng(t)=E\{J_{AB}\}=E\{J_{AC}\}=E\{J_{BC}\}$ and also $Ng(t)(1-g(t))=Var(J_{AB})=Var(J_{AC})=Var(J_{BC})=\sigma^2$ we have that the distribution of $J_x$ is normal with mean $E\{J_{AB}\}$ and variance $(\sigma^2+Cov)/2$.

Similarly, $J_x-J_{AC}$ is normal, with mean 0 and variance $Var(J_x)+ Var(J_{AC}) -2Cov(J_x,J_{AC})$

Now,

$$Cov(J_{\bar{X}}, J_{AC}) =Cov\left(\frac{J_{AB}+J_{BC}}{2}, J_{AC}\right)$$

$$=E\left[\frac{(J_{AB}+J_{BC})\,J_{AC}}{2}\right] -E\left[\frac{(J_{AB}+J_{BC})}{2}\right] E[J_{AC}]$$

$$=E\left[\frac{(J_{AB}J_{AC})}{2}+\frac{(J_{BC}J_{AC})}{2}\right] -\frac{E[J_{AB}]\,E[J_{AC}]}{2}-\frac{E[J_{BC}]\,E[J_{AC}]}{2}$$

$$= \frac{Cov(J_{AB}, J_{AC})}{2} + \frac{Cov(J_{BC}, J_{AC})}{2}$$

= Cov

Thus,

$Var\ (J_x - J_{AC}) = Var(J_x) + Var(J_{AC}) - 2\ Cov(J_x, J_{AC})$

$= (\sigma^2 + Cov)/2 + \sigma^2 - 2\ Cov$

$= 3(\sigma^2 - Cov)/2$

hence

$$d = \frac{(J_{\bar{X}} - J_{AC})}{\sqrt{3\,(\sigma^2 - Cov)\,/2}}$$

is normal standard.

**Appendix G**

Let $P(\text{Wrong tree} \mid H_0 \text{ false}) = P_{H1}(Wt)$

Given $H_1$ true, we let A and C be the species with shorter time to divergence, $t_1$, and let $t_0$ the time from divergence of pairs AB and BC, $t_1 < t_0$. Without loss of generality, let $J_{AB}$ be the smallest of the three random variables. We have:

$P_{H1}(Wt) = P(J_{AC}$ is not the smallest and $H_0$ is rejected $\mid H_1$ true)

Using conditonal probability:

$P_{H1}(Wt) = P(J_{AC}$ is not the smallest) $P(H_0$ is rejected given that $J_{AC}$ is not the smallest $\mid H_1$ true)

The above probability depends on $t_0$ and $t_1$, but it is possible to find an upper limit. Note that the probability that $J_{AC}$ will not be the smallest increases as $t_1$ approaches $t_0$, since $E\{J_{AC}\}$ is an increasing function of time, on the other hand, for $t_0$ fixed, the probability that $H_0$ will be rejected given that $J_{AC}$ is not the smallest depends on the size of the difference $J_x - J_{AB}$, that is, on:

$$\frac{1}{2}J_{BC} + \frac{1}{2}J_{AC} - J_{AB}$$

The above distance can be maximized by letting either $J_{AC} - J_{AB}$

or $J_{BC}-J_{AB}$ be as big as possible, nevertheless, since $J_{AB}$ and $J_{BC}$ are two samples of Bin(N, $g(t_0)$), for $t_0$ fixed, we can only manipulate the distance $J_{AC}-J_{AB}$. Thus, to maximize $J_x-J_{AB}$ it is neccesary that $t_1$ approaches $t_0$.

We conclude that both, $P(J_{AC}$ is not the smallest) and $P(H_0$ is rejected given that $J_{AC}$ is not the smallest$|H_1$ true) are maximized when the pairs AC, AB and BC have very similar time to divergence. We let $t_0=t_1$, thus:

$$\max P_{H1}(Wt) = P(J_{AC} \text{ is not the smallest} | t_0=t_1)\ P(H_0 \text{ is rejected given that } J_{AC} \text{ is not the smallest } | t_0=t_1)$$
$$= P(J_{AC} \text{ is not the smallest} | H_0 \text{ true})\ P(H_0 \text{ is rejected given that } J_{AC} \text{ is not the smallest } | H_0 \text{ true})$$

Since

$P(J_{AC}$ is the smallest$|H_0$ true) + $P(J_{AC}$ is not the smallest$|H_0$ true) = 1

we have, by the total probability law:

$P(\text{Reject } H_0 | H_0 \text{ true}) =$
$P(J_{AC} \text{ is the smallest} | H_0 \text{ true})\ P(\text{Reject } H_0 | J_{AC} \text{ the smallest } H_0 \text{ true}) + P(J_{AC} \text{ not the smallest} | H_0 \text{ true})\ P(\text{Reject } H_0 | J_{AC} \text{ not the smallest}, H_0 \text{ true})$

Since the maximum of $P_{H1}(Wt)$ involves the second term of the right side in the last expression, this can be rewritten as:

$P(\text{Reject } H_0 | H_0 \text{ true}) =$
$P(J_{AC} \text{ is the smallest} | H_0 \text{ true})\ P(\text{Reject } H_0 | J_{AC} \text{ the smallest } H_0 \text{ true}) + P(\text{Wrong tree} | H_0 \text{ false})$

It follows that for a level $\alpha$ test:

$P(\text{Wrong tree} | H_0 \text{ false}) < P(\text{Reject } H_0 | H_0 \text{ true}) < \alpha$

## REFERENCES

Brown, W.M., E.M. Prager, A. Wang and A.C. Wilson. 1982. Mitochondrial DNA sequences of primates: Tempo and mode evolution. J. Mol. Evol. 18:225-239.

Casella, G. and R.L. Berger. 1990. Statistical Inference. Waldswort Inc.

Felsenstein, J. 1983. Methods for inferring phylogenies: A statistical view. In Numerical Taxonomy, J. Felsenstein (ed.). Spring-Verlag, Berlin.

Jukes, T.H. and C.R. Cantor. 1969. Evolution in protein molecules. In Mammalian Protein Metabolism, H.N. Munro (ed.). Academic Press, New York.

Patel, J.K.,C.H. Kapadia and D.B. Owen, 1976. Handbook of Statistical Distributions. Marcel Dekker, Inc. New York.

Tsokos, C. 1972. Probability distributions: An introduction to probability theory with applications. Duxbury Press. pp. 540-541.

Weir, B.S. 1990. Genetic data analysis. Sinauer Associates, Inc. Sunderland, Massachusetts.